|  |
| --- |
| > setwd('C:/Rdirectory/data\_mining/data\_mining\_covid');  > covid\_train = read.csv("covid\_train.csv", header=T);  > covid\_train$sex = as.factor(covid\_train$sex);  > covid\_train$patient\_type = as.factor(covid\_train$patient\_type);  > covid\_train$pneumonia = as.factor(covid\_train$pneumonia);  > covid\_train$age = as.numeric(covid\_train$age);  > covid\_train$diabetes = as.factor(covid\_train$diabetes);  > covid\_train$copd = as.factor(covid\_train$copd);  > covid\_train$asthma = as.factor(covid\_train$asthma);  > covid\_train$inmsupr = as.factor(covid\_train$inmsupr);  > covid\_train$hypertension = as.factor(covid\_train$hypertension);  > covid\_train$other\_disease = as.factor(covid\_train$other\_disease);  > covid\_train$cardiovascular = as.factor(covid\_train$cardiovascular);  > covid\_train$obesity = as.factor(covid\_train$obesity);  > covid\_train$renal\_chronic = as.factor(covid\_train$renal\_chronic);  > covid\_train$tobacco = as.factor(covid\_train$tobacco);  > covid\_train$contact\_other\_covid = as.factor(covid\_train$contact\_other\_covid);  > covid\_train$is\_dead = as.factor(covid\_train$is\_dead);  > str(covid\_train);  'data.frame': 32356 obs. of 16 variables:  $ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...  $ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...  $ pneumonia : Factor w/ 2 levels "1","2": 2 2 1 1 1 2 1 1 1 2 ...  $ age : num 25 52 51 67 59 52 54 78 80 40 ...  $ diabetes : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 2 2 1 2 ...  $ copd : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ hypertension : Factor w/ 2 levels "1","2": 2 2 2 1 2 1 2 1 1 2 ...  $ other\_disease : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 2 2 2 2 ...  $ obesity : Factor w/ 2 levels "1","2": 2 1 2 1 2 2 2 1 1 2 ...  $ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ tobacco : Factor w/ 2 levels "1","2": 2 1 2 2 2 2 2 2 2 2 ...  $ contact\_other\_covid: Factor w/ 2 levels "1","2": 1 1 1 1 1 2 2 1 1 2 ...  $ is\_dead : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 1 2 2 ...  > head(covid\_train);  sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco  1 1 2 2 25 2 2 2 2 2 2 2 2 2 2  2 1 2 2 52 2 2 2 2 2 2 2 1 2 1  3 1 2 1 51 2 2 2 2 2 2 2 2 2 2  4 1 2 1 67 1 2 2 2 1 2 2 1 2 2  5 1 2 1 59 1 2 2 2 2 2 2 2 2 2  6 1 2 2 52 1 2 2 2 1 2 1 2 2 2  contact\_other\_covid is\_dead  1 1 2  2 1 2  3 1 2  4 1 2  5 1 2  6 2 2  > covid\_test = read.csv("covid\_test.csv", header=T);  > covid\_test$sex = as.factor(covid\_test$sex);  > covid\_test$patient\_type = as.factor(covid\_test$patient\_type);  > covid\_test$pneumonia = as.factor(covid\_test$pneumonia);  > covid\_test$age = as.numeric(covid\_test$age);  > covid\_test$diabetes = as.factor(covid\_test$diabetes);  > covid\_test$copd = as.factor(covid\_test$copd);  > covid\_test$asthma = as.factor(covid\_test$asthma);  > covid\_test$inmsupr = as.factor(covid\_test$inmsupr);  > covid\_test$hypertension = as.factor(covid\_test$hypertension);  > covid\_test$other\_disease = as.factor(covid\_test$other\_disease);  > covid\_test$cardiovascular = as.factor(covid\_test$cardiovascular);  > covid\_test$obesity = as.factor(covid\_test$obesity);  > covid\_test$renal\_chronic = as.factor(covid\_test$renal\_chronic);  > covid\_test$tobacco = as.factor(covid\_test$tobacco);  > covid\_test$contact\_other\_covid = as.factor(covid\_test$contact\_other\_covid);  > covid\_test$is\_dead = as.factor(covid\_test$is\_dead);  > str(covid\_test);  'data.frame': 5920 obs. of 16 variables:  $ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...  $ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...  $ pneumonia : Factor w/ 2 levels "1","2": 2 2 1 1 1 1 2 2 2 1 ...  $ age : num 52 36 0 85 75 19 46 60 62 39 ...  $ diabetes : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 1 2 ...  $ copd : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 2 2 2 ...  $ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ hypertension : Factor w/ 2 levels "1","2": 2 2 2 1 1 2 2 2 2 1 ...  $ other\_disease : Factor w/ 2 levels "1","2": 1 2 2 2 2 2 2 2 2 2 ...  $ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 2 2 2 ...  $ obesity : Factor w/ 2 levels "1","2": 2 2 2 1 2 2 2 2 2 2 ...  $ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ tobacco : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ contact\_other\_covid: Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 2 2 2 ...  $ is\_dead : Factor w/ 2 levels "1","2": 2 2 1 2 2 2 2 2 2 2 ...  > head(covid\_test);  sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco  1 1 2 2 52 2 2 2 2 2 1 2 2 2 2  2 1 2 2 36 2 2 2 2 2 2 2 2 2 2  3 1 2 1 0 2 2 2 2 2 2 2 2 2 2  4 1 2 1 85 2 1 2 2 1 2 1 1 2 2  5 1 2 1 75 2 2 2 2 1 2 2 2 2 2  6 1 2 1 19 2 2 2 2 2 2 2 2 2 2  contact\_other\_covid is\_dead  1 2 2  2 2 2  3 2 1  4 2 2  5 2 2  6 2 2  > covid\_dead\_train = read.csv("covid\_dead\_train.csv", header=T);  > covid\_dead\_train$sex = as.factor(covid\_dead\_train$sex);  > covid\_dead\_train$patient\_type = as.factor(covid\_dead\_train$patient\_type);  > covid\_dead\_train$pneumonia = as.factor(covid\_dead\_train$pneumonia);  > covid\_dead\_train$age = as.numeric(covid\_dead\_train$age);  > covid\_dead\_train$diabetes = as.factor(covid\_dead\_train$diabetes);  > covid\_dead\_train$copd = as.factor(covid\_dead\_train$copd);  > covid\_dead\_train$asthma = as.factor(covid\_dead\_train$asthma);  > covid\_dead\_train$inmsupr = as.factor(covid\_dead\_train$inmsupr);  > covid\_dead\_train$hypertension = as.factor(covid\_dead\_train$hypertension);  > covid\_dead\_train$other\_disease = as.factor(covid\_dead\_train$other\_disease);  > covid\_dead\_train$cardiovascular = as.factor(covid\_dead\_train$cardiovascular);  > covid\_dead\_train$obesity = as.factor(covid\_dead\_train$obesity);  > covid\_dead\_train$renal\_chronic = as.factor(covid\_dead\_train$renal\_chronic);  > covid\_dead\_train$tobacco = as.factor(covid\_dead\_train$tobacco);  > covid\_dead\_train$contact\_other\_covid = as.factor(covid\_dead\_train$contact\_other\_covid);  > covid\_dead\_train$day\_cnt = as.numeric(covid\_dead\_train$day\_cnt);  > str(covid\_dead\_train);  'data.frame': 3600 obs. of 16 variables:  $ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...  $ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...  $ pneumonia : Factor w/ 2 levels "1","2": 1 1 2 1 1 1 1 1 1 2 ...  $ age : num 78 65 58 78 92 70 71 65 56 72 ...  $ diabetes : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 1 2 2 ...  $ copd : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 2 2 2 ...  $ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ hypertension : Factor w/ 2 levels "1","2": 1 2 2 2 1 2 1 1 2 1 ...  $ other\_disease : Factor w/ 2 levels "1","2": 2 2 1 2 2 2 2 2 2 2 ...  $ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ obesity : Factor w/ 2 levels "1","2": 1 1 2 1 2 1 1 1 2 2 ...  $ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 1 ...  $ tobacco : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 1 2 2 2 ...  $ contact\_other\_covid: Factor w/ 2 levels "1","2": 1 2 1 1 2 2 1 2 2 2 ...  $ day\_cnt : num 10 7 1 6 3 16 21 14 30 20 ...  > head(covid\_dead\_train);  sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco  1 1 2 1 78 2 2 2 2 1 2 2 1 2 2  2 1 2 1 65 2 2 2 2 2 2 2 1 2 2  3 1 2 2 58 2 2 2 2 2 1 2 2 2 2  4 1 2 1 78 2 2 2 2 2 2 2 1 2 2  5 1 2 1 92 2 2 2 2 1 2 2 2 2 2  6 1 2 1 70 2 2 2 2 2 2 2 1 2 2  contact\_other\_covid day\_cnt  1 1 10  2 2 7  3 1 1  4 1 6  5 2 3  6 2 16  > covid\_dead\_test = read.csv("covid\_dead\_test.csv", header=T);  > covid\_dead\_test$sex = as.factor(covid\_dead\_test$sex);  > covid\_dead\_test$patient\_type = as.factor(covid\_dead\_test$patient\_type);  > covid\_dead\_test$pneumonia = as.factor(covid\_dead\_test$pneumonia);  > covid\_dead\_test$age = as.numeric(covid\_dead\_test$age);  > covid\_dead\_test$diabetes = as.factor(covid\_dead\_test$diabetes);  > covid\_dead\_test$copd = as.factor(covid\_dead\_test$copd);  > covid\_dead\_test$asthma = as.factor(covid\_dead\_test$asthma);  > covid\_dead\_test$inmsupr = as.factor(covid\_dead\_test$inmsupr);  > covid\_dead\_test$hypertension = as.factor(covid\_dead\_test$hypertension);  > covid\_dead\_test$other\_disease = as.factor(covid\_dead\_test$other\_disease);  > covid\_dead\_test$cardiovascular = as.factor(covid\_dead\_test$cardiovascular);  > covid\_dead\_test$obesity = as.factor(covid\_dead\_test$obesity);  > covid\_dead\_test$renal\_chronic = as.factor(covid\_dead\_test$renal\_chronic);  > covid\_dead\_test$tobacco = as.factor(covid\_dead\_test$tobacco);  > covid\_dead\_test$contact\_other\_covid = as.factor(covid\_dead\_test$contact\_other\_covid);  > covid\_dead\_test$day\_cnt = as.numeric(covid\_dead\_test$day\_cnt);  > str(covid\_dead\_test);  'data.frame': 420 obs. of 16 variables:  $ sex : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...  $ patient\_type : Factor w/ 1 level "2": 1 1 1 1 1 1 1 1 1 1 ...  $ pneumonia : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...  $ age : num 66 68 49 61 47 56 68 72 78 64 ...  $ diabetes : Factor w/ 2 levels "1","2": 1 2 2 1 1 2 1 1 1 1 ...  $ copd : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 1 2 2 ...  $ asthma : Factor w/ 2 levels "1","2": 2 2 2 2 1 2 2 2 2 2 ...  $ inmsupr : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ hypertension : Factor w/ 2 levels "1","2": 1 2 2 1 2 2 1 1 2 2 ...  $ other\_disease : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ cardiovascular : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ obesity : Factor w/ 2 levels "1","2": 1 2 1 2 2 2 1 1 1 2 ...  $ renal\_chronic : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ tobacco : Factor w/ 2 levels "1","2": 2 2 2 2 2 2 2 2 2 2 ...  $ contact\_other\_covid: Factor w/ 2 levels "1","2": 1 2 2 2 2 2 2 2 2 2 ...  $ day\_cnt : num 3 5 16 4 18 10 9 6 6 8 ...  > head(covid\_dead\_test);  sex patient\_type pneumonia age diabetes copd asthma inmsupr hypertension other\_disease cardiovascular obesity renal\_chronic tobacco  1 1 2 1 66 1 2 2 2 1 2 2 1 2 2  2 1 2 1 68 2 2 2 2 2 2 2 2 2 2  3 1 2 1 49 2 2 2 2 2 2 2 1 2 2  4 1 2 1 61 1 2 2 2 1 2 2 2 2 2  5 1 2 1 47 1 2 1 2 2 2 2 2 2 2  6 1 2 1 56 2 2 2 2 2 2 2 2 2 2  contact\_other\_covid day\_cnt  1 1 3  2 2 5  3 2 16  4 2 4  5 2 18  6 2 10  > library(partykit);  필요한 패키지를 로딩중입니다: grid  필요한 패키지를 로딩중입니다: libcoin  필요한 패키지를 로딩중입니다: mvtnorm  > library(CHAID);  > covid\_train = covid\_train[,!names(covid\_train) %in% c("age")];  > covid\_test = covid\_test[,!names(covid\_test) %in% c("age")];  > covid\_dead\_train = covid\_dead\_train[,!names(covid\_dead\_train) %in% c("age")];  > covid\_dead\_test = covid\_dead\_test[,!names(covid\_dead\_test) %in% c("age")];  > chaidFit = chaid(is\_dead~., data = covid\_train, control = chaid\_control(minsplit = 5000, minprob = 0.4));  > plot(chaidFit);  > print(chaidFit);  Model formula:  is\_dead ~ sex + patient\_type + pneumonia + diabetes + copd +  asthma + inmsupr + hypertension + other\_disease + cardiovascular +  obesity + renal\_chronic + tobacco + contact\_other\_covid  Fitted party:  [1] root  | [2] pneumonia in 1  | | [3] hypertension in 1  | | | [4] diabetes in 1  | | | | [5] renal\_chronic in 1: 1 (n = 849, err = 26.1%)  | | | | [6] renal\_chronic in 2: 1 (n = 4971, err = 31.7%)  | | | [7] diabetes in 2: 1 (n = 4507, err = 35.3%)  | | [8] hypertension in 2  | | | [9] diabetes in 1: 1 (n = 3544, err = 35.8%)  | | | [10] diabetes in 2  | | | | [11] contact\_other\_covid in 1: 2 (n = 2779, err = 42.0%)  | | | | [12] contact\_other\_covid in 2  | | | | | [13] cardiovascular in 1: 1 (n = 240, err = 29.6%)  | | | | | [14] cardiovascular in 2  | | | | | | [15] copd in 1: 1 (n = 259, err = 33.2%)  | | | | | | [16] copd in 2  | | | | | | | [17] obesity in 1: 1 (n = 1489, err = 43.6%)  | | | | | | | [18] obesity in 2  | | | | | | | | [19] renal\_chronic in 1: 1 (n = 141, err = 36.2%)  | | | | | | | | [20] renal\_chronic in 2: 1 (n = 5603, err = 48.3%)  | [21] pneumonia in 2  | | [22] hypertension in 1: 2 (n = 2123, err = 34.9%)  | | [23] hypertension in 2  | | | [24] contact\_other\_covid in 1: 2 (n = 2182, err = 7.6%)  | | | [25] contact\_other\_covid in 2: 2 (n = 3669, err = 19.7%)  Number of inner nodes: 12  Number of terminal nodes: 13  > prediction = predict(chaidFit, newdata=covid\_test[]);  > # prediction = predict(chaidFit, newdata=covid\_test[], type="prob");  > summary(prediction);  1 2  3321 2599 |
| > comparison=cbind(covid\_test,prediction);  comparison=as.data.frame(comparison);  print(paste("test 건수 : ",nrow(covid\_test)));  predictCorrect = comparison[comparison$is\_dead == comparison$prediction,];  print(paste("사망여부 예측성공 건수 : ", nrow(predictCorrect))); #3579  print(paste("사망여부 예측 정확도 : " ,nrow(predictCorrect)/nrow(covid\_test))); # 60.4% |

